Crowd Emotion Recognition

Jennifer Swaminathan
University of Neuchatel
Neuchatel, Switzerland
Email: jennifer.swaminathan@unine.ch

Marine Capallera
University of Fribourg
Fribourg, Switzerland
Email: marine.capallera@hefr.ch

***Abstract*—Crowd emotion recognition is of great significance in various fields especially in maintaining public safety. The behavior of the crowd can be predicted by monitoring emotions to prevent stampedes and riots in large events. This literature review studies recent crowd emotion recognition models and compares the different criteria used to design these models. We observe a pipeline is followed to design this model which includes the input modalities, feature extraction, algorithms to detect the emotions, and the output component to predict the behavior of the crowd. In the end, we propose tools that can be used to induce these emotions in digital experiences.**

I. Introduction

Crowd emotion recognition is a rapidly growing field, that uses machine learning algorithms to analyze the emotion and take appropriate actions. One of the main advantages of crowd emotion recognition is its ability to obtain valuable insights into the emotional state of the crowd.

Detecting the emotion of the crowd, can be useful in various contexts. For example, Baig et al in [1], suggests a scenario, where in a music concert, with video surveillance, we can detect the emotions felt by the crowd. By keeping track of the emotions, we can predict the behavior of the crowd and before a behavior change of 'Panic' occurs, meaning a stampede, the authorities can be warned to take quick action. Alternatively, when the crowd's emotions are showing 'bored' by using external simulation, the mood of the participants can be improved.

There are different methods that can be used to detect emotions in individuals. For example, using facial expressions, the emotions felt by the individual can be detected. Voice expressions can be used to detect emotions, through Natural Language Processing. Text messages sent on social media can also express emotions. Finally, body posture of an individual can also be used for emotion detection.

But these are not easily translated when we talk about a crowd. In the case of individuals, we could use wearable technology to analyze the body signals, whereas when we consider the crowd as a whole, this is not feasible. In order to analyze the crowd, we need monitoring systems. The common crowd-monitoring system is video surveillance. By monitoring the behavior of the crowd in real time, we can prevent and intervene when there is an issue. With current advanced technology, facial recognition, and high-resolution cameras, these emotions can be easily classified. The other modality to monitor crowd emotions is sound. Fasisal et al. in [3] use the background emotions felt by the crowd as a whole and not the speech which was given by the main actor.

The shortcomings of these crowd emotion recognition models are the amount of data processing needed to detect emotions. Detecting the emotions in the crowd through facial recognition is computationally expensive. There could be view points ignored by the surveillance camera. Finally, privacy concerns are also an issue as the technology is monitoring individuals in public places.

This literature review presents the state-of-the-art techniques used in crowd emotion detection and explores the possibility of tools available to induce these emotions to have a sensory experience. We answer the following research questions:

Q1: What is the pipeline followed to build crowd emotion recognition?

Q2: In which context is crowd emotion recognition utilized?

Q3: Can crowd emotions be induced during digital experiences?

The structure of the literature review is as follows. Section 2 presents the definitions of terms used in crowd emotion recognition models. Section 3 presents the analysis conducted. In Section 3.A, we present, the methodology used to answer our research questions. Section 3.B provides the results we obtained which include a comparison of the criteria used in different models, and a synthesis table of the different models. Section 4, discusses affective haptic devices as a tool to induce emotions and enhance digital experience. Section 5, provides a discussion of the review. Section 6, contains the conclusion and possible future work.

II. Definitions

In this section, we clarify important terms related to crowd emotion recognition. According to Sanchez et al. in [14], a crowd is defined as "a unique large group of individuals sharing a common physical location." The density and common location of the group are key factors in determining whether it can be considered a crowd. For instance, a group of 20 people in an elevator can be considered a crowd, but the same number of individuals in a stadium would not. It is crucial to consider crowd density and physical location in the definition of a crowd.

But using only the density of the crowd, crowds emotion cannot be identified. Research has been done on anomaly detection from video clips by considering only crowd motion and density, but it was insufficient to predict the behavior. For

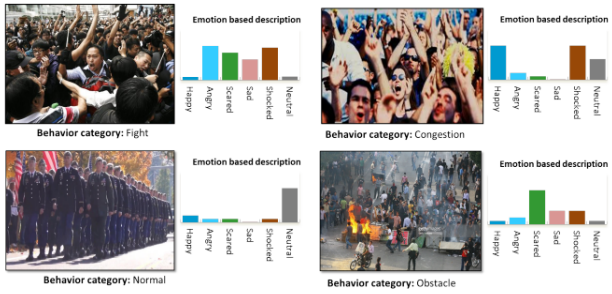

Fig. 1. Emotion based crowd representation extracted from [13]

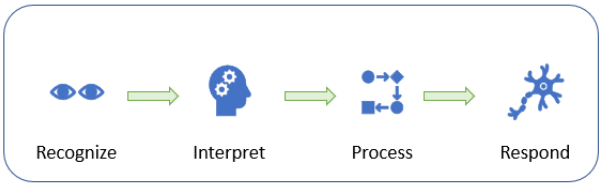

Fig. 2. Affective Computing in Crowd Emotion Recognition

example, crowd activities that appear to be visually similar as in Fig 1, display different emotions. All four images have different levels of emotions measured even though they look similar with high density. Hence, Rabiee et al in [13], suggest using crowd emotions, in order to detect anomalies.

Emotions are triggered by internal and external factors and are characterized by physiological, behavioral, and cognitive changes. Emotions can be represented on a two-dimensional plane, with valence on the horizontal axis and arousal on the vertical axis. Valence can be positive (associated with happiness, and love) and negative (associated with sadness and anger). Arousal refers to the level of intensity of emotion. High arousal is excitement, whereas low arousal is calm.

Affective computing is a field of study focused on creating systems and devices capable of recognizing, interpreting, processing, and responding to human emotions. This technology is used in crowd emotion recognition to detect the emotions of a crowd and respond accordingly, making it valuable in event management. By monitoring the emotional state of a crowd, affective computing helps identify signs of stress, anxiety, or aggression and take appropriate action.

III. Analysis

A. Methodology

We started our literature review on crowd emotion detection by searching for research papers on Semantic Scholar. Following were the queries we used, ”crowd emotion detection”, and ”crowd emotion analysis events”. We obtained various papers as our result. We filtered our search to the past 10 years and the papers with more citations. We removed papers that were based on sentiment analysis on text and selected papers that used video or sound to detect emotions and chose 12 papers. From the selected papers, we performed an analysis based on the following criteria: context, the algorithm used to obtain the crowd emotions, input form, feature extraction from the input, the output obtained from the algorithm, evaluations performed, and the number of emotions detected by each model.

To answer our Q3, we found out there are tools we could use to simulate these crowd emotions. For example, one of the methods is fully immersive virtual reality, where the user can utilise VR goggles, and be placed in the virtual environment to experience crowd emotions. 5D effects is another method, where the participants are placed in chairs along with external devices like 3D glasses. Along with seat movements and wind, sound effects we could simulate crowd emotions. The other tool that is used to induce emotions is Affective Haptic devices(AHD). We decided to explore these AHD tools and searched semantic scholar for affective haptic devices to induce crowd emotions, but we could not find papers related to crowd emotions, so used ’affective haptic devices induce emotions’ and selected a few research papers, to propose an idea to stimulate crowd emotions using affective computing.

B. Results

In this subsection, we present our results from the analysis performed. We present a synthesis table of the models we studied, in Table I. This synthesis table compares the important criteria of the models studied namely, the aim of the models, the context where these models are used, the emotion detection algorithm used, input modalities, feature extraction, output results from the model, evaluation metric, and the number of emotions they can classify.

1) Context: To apply crowd emotion detection, context is key. In Martini et al. [8] crowd emotion recognition is used to light up the Christmas tree present in a shopping mall present in Milan as shown in Fig 3. This increases the shopping experience of the customers. Sanchez et al in[14], suggest using crowd emotion detection during Covid19 pandemic, in order to detect crowds assembled in public areas. In [17], [16], suggests using crowd emotion recognition models to avoid violence. By studying the emotions of the crowd, we can prevent violent behavior of the crowd, by predicting the behavior of the crowd. Zhang et al. in [18], uses them to detect any anomaly, like a truck on pedestrian roads, or sharp objects detection. In [10], uses crowd emotion recognition in indoor pubs, events, to classify the emotion of a group or crowd watching a match on a screen.

In [3], [4], suggests using crowd emotion detection in crowded events like in stadiums, to survey the crowd to pay attention to the emotions felt by the crowd to predict any violence that could occur and alert the authorities. Faisal et al. [3], suggest using CER in movie theatres to provide information on which parts are enjoyed by the audience and use them to find the quality of the movie and use these parts for advertisements to increase revenue.

TABLE I
SYNTHESIS

	Causal Spatio-temporal Structure [17]	**Evaluation based on fuzzy inference [18]**	**Crowd Emotion to light up smart Christmas Tree[8]**	**ESLCE: Emotional Sounds[3]**	**CE sounds: spectrogram-based analysis using CNN[4]**	**NVPF to group-level emotion recognition [12]**	**2D ConvNets[16]**	**CED Bio-inspired [1]**	**Edge based super-imposed CER[10]**	**Crowd Abnormal Behavior Detection[7]**	**DL Models Combined with CE Models[6]**	**Emotion-Based Crowd Represen-tation [13]**
Year	2022	2021	2019	2021	2019	2022	2020	2014	2016	2020	2020	2016
Aim	instead of traditional DL, causal inference with RNN	to measure the crowd emotion-using fuzzy inference	Improve re-tail experi-ence	Curate a dataset for detecting crowd emotion with sound	CE from speech and sounds clips	Multi level emotion classifica-tion	CE using 2D CNN	detection of crowd emotion and simulate it.	spontaneous detection of emotion even with occlusion	using motion and emotion to detect anomaly	OCC with DL methods to detect abnormality.	Use crowd motion and emotion to find abnormality.
Context	Avoid violence	Emotion detection in crowd	Light up Christmas Tree	CE in real life Events	Events and voilence	Emotion in a crowd with different groups	Peaceful Protests	Anomaly detection	Indoor crowd events	abnormality	abnormality	abnormality
Algorithm	Cs-RNN with ITE	Fuzzy rules	CNN	RandomForest classifica-tion	visual trans-fer learning	CNN, NVPF, TNVPF	2D-CNN	bayesian networks, bio-inspired memory	edge detec-tion , SVM classifier	dual channel CNN and SVM	CNN, OCC emotion model	SVM classi-fier
Input	Video	Video	photos	sounds	sounds	video and images	image	Video	Video	Video	Video	Video
Feature extraction	crowd density, crowd enthalpy, magnitude variance, confusion index	crowd density, crowd enthalpy, magnitude variance, confusion index	Sentiment extraction from photos	Loudness, audio-specific features	frequency-amplitude features	individual facial detection (bottom-up)	facial features using DL	Emotion and motion	edges and motion from crowd	Emotion and motion	crowd entropy, emotion from OCC model	low level features from dense trajectories
Output	classification of emotion	valence, arousal score	light the tree	classification of emotion	classification of emotion	classification of multiple level of emotions	classification of emotion	classification of emotion	classification of emotion	classification of the behavior	classification of behavior	classification of behavior
Results	CS-RNN -performed better	MRE, MAE are relatively low	accuracy good for happy and neutral	F1 score Disapproval is poorly identified	random train-test better than manual selection	Robust and effective	Good accu-racy is ob-tained	Accuracy is good (except for Herding)	Accuracy and recall, happy is the best, anger is the least	AUC,ERR and accuracy	Accuracy better with emotion AUC and EER	Good accuracy when emotion is considered
Dataset	MED	own dataset	own dataset	own dataset	own	own-GEVC	self curated	simulated	own	UCSD	MED, UCSD	own
No of emo-tions	6	12	4	3	3	3	5	2	7	7	do not men-tion	6

Fig. 3. Christmas Tree illuminated extracted from [8]

2) Input: To detect the crowd's emotions, we need data. These data can be in the form of video, images, or sounds.

In [4], [3], uses crowd speech and sound in mass events. With the sound, context is key to analyzing the sound. For example, a crowd could shout in anger when an opposition team scores or due to fear induced by a panic situation. The crowd sounds consist of speech as well as clatter based, that is booing, whistling, and shaking objects. Furthermore, in these sound clips, they filter out the main speaker's speech and only use the noise from the crowd. Since the speaker's speech only detects the emotion of the speaker. So the input consists of crowd speech and clatter. Martini et al in [8], use images posted by various participants to light the Christmas tree. Tripathi et al in [16], also use images as input to extract the emotions found in the crowd. In [17], [18], [13], the authors use videos as their main source of input to classify the behavior of the crowd. Quach et al. in [12], use input modalities that can be either images, and videos.

3) Feature Extraction: The subsequent step after obtaining the input is to extract the features from these data, in order to classify the emotion. From the available data, features are extracted to classify them into emotions. In cite [17], [18], from the videos, enthaply of crowd, Magnitude Variance, Confusion Index, and Crowd Density are obtained after removal of background and foreground fields. Higher crowd density means smaller space a single person owns and increases depression. Enthalpy describes the state of the crowd system. When there are more dramatic movements in the crowd, enthalpy increases. Magnitude variance describes the consistency of the crowd movement velocity. Greater the variance, the more intense the emotion. Confusion index is used to describe the consistency of the crowd movement direction. For example, chaotic movement leads to unpleasant emotions and increases the confusion index. In [17], the pipeline used to extract these features is shown in Fig 4. In [8], the facial features are extracted from the images to classify the emotion using CNN. In [3], [4], loudness, frequency, and other audio-specific features are extracted. To extract the features[3], they convert the stereo signal to mono signal. Next these signals are normalised and labeled based on the emotions expressed. They use a total of 34 features.

In Tripathi et al. [16], facial features are extracted from the videos using Deep learning techniques.

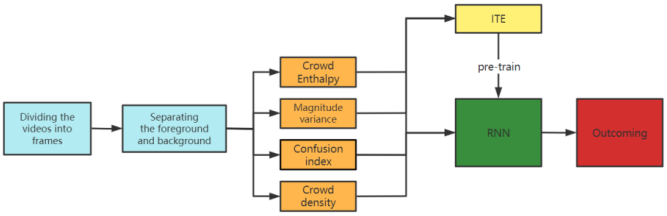

Fig. 4. Crowd Emotion Recognition based on Causal Spatiotemporal Structure extracted from [17]

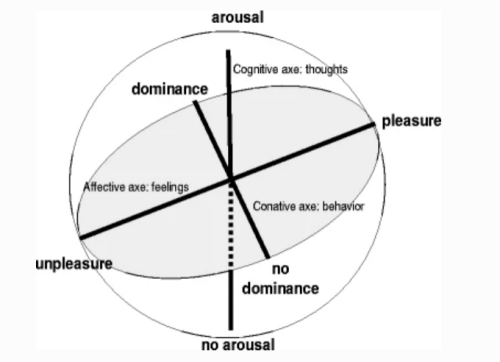

Fig. 5. 3-dimensional PAD model extracted from [2]

Rabiee et al in [13], uses density trajectories as features along with the emotions which are annotated in their training dataset, in order to classify the behavior of the crowd.

Li et al. in [6], extract crowd entropy by using the microscopic state of the segmented video. When the entropy value calculated is higher than a predefined threshold, the crowd emotion state is calculated using OCC emotion model. OCC model (Ortony, Clore and Collins) mapped emotions according to PAD(Pleasure, Arousal, Dominance) emotion model. In Bakker et al. [2], suggests the provided 3D PAD model, as shown in Fig 5 considers all human emotions compared to the 2D model, valence, and arousal. So [6], suggests that when the crowd's emotion exceeds a threshold they communicate their emotions via their behavior, hence we can predict the next state. These features extracted from the videos are provided as input to the CNN classifier.

Li et al in [7], use motion and emotion features to classify the behavior. In order to extract the motion features, the trajectories of the crowd are tracked. This is achieved by using dual-channel CNN. Channel 1, is used to extract the motion features, and Channel 2 to extract the emotion probabilities.

In Patwardhan et al. [10], with the given video input, 8 frames per second are used to extract features. They follow the flowchart in Fig 6 to extract the features. Initially, they extract edges from the frame, which is obtained by extracting high-intensity changes. In Fig 7, we can see how the edges are extracted. After extracting the edges, an optimized grid is placed on each frame and the temporal, kinetic, and motion pattern features are extracted to obtain 400 features.

Quach et al. [12], proposes a high-performance and low-

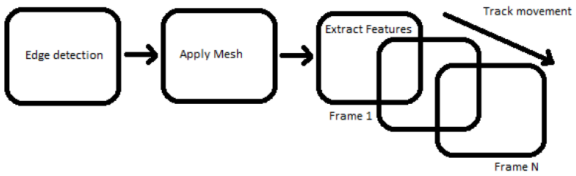

Fig. 6. Flow chat Feature extraction extracted from [10]

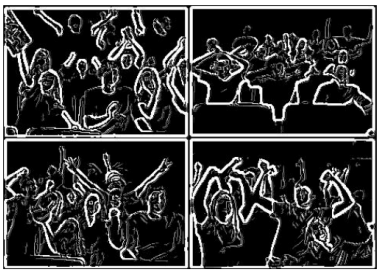

Fig. 7. Edge detection extracted from [10]

TABLE II
EMOTION TO BEHAVIOR EXTRACTED FROM [7]

Emotions	Behaviors
Angry	Fight
Happy	Cheerful
Excited	Cheerful
Scared	Panic
Sad	Congestion
Neutral	Normal
Nothing	Abnormal objects

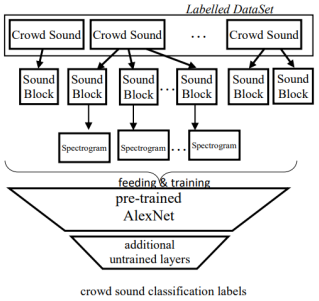

Fig. 8. FLowchart of Heterogeneous transfer learning extracted from [4]

cost deep network for extracting facial expressions of individuals, which are then provided to the Non-Volume preserving fusion(NVPF).

4) Algorithm: In this subsection, we compare the different algorithms used by the models to detect crowd emotion.

In [8], facial emotion recognition is performed using CNN. In CNN, the input image is convolved through a collection of filters to produce a map of features, and the last layer of the network classifies the facial expression.

In [16], 2D CNN is used to extract the features from an image. It has an input layer of 300X300X3 dimensions. From these input layer, 2D Convolution network along with rectifier linear unit activation function is deployed to classify into 5 emotions. In [13], to classify the emotions from a given video, SVM classifier is used. From the video clips, features are extracted. By considering only low-level features like crowd density, the accuracy of the model was poor. The reason for this is that congestion can be a result of 'happy' or 'angry'. Hence they consider the emotions in the video, along with the features to form a latent SVM classifier, which provides a more efficient model.

Li et al. [7], the output from the dual CNN channel with features, is provided to a multi-class SVM, to classify the emotions into the behavior as shown in Table II.

In [10], after extraction of the features using edge detection and applying mesh, 400 features were obtained. Using dimensionality reduction 23 discriminatory feature vectors were selected. These videos were annotated by 3 observers to classify them as happy, angry, surprised, sad, disgust, fear, and neutral. The feature vectors were used to train the SVM classifier, using 10-fold cross validation.

In Baig et al. [1] to predict the emotions of the crowd before it occurs, they memorize all interactions that occur in a given context. A bio-inspired algorithm is considered. Here, the interactions that occur in a situation are memorized, for example, the reaction of a person, due to the action of a crowd, and the situation in which this emotion is felt is noted. Different events are stored in the memory. An event is defined by the changes in proto(person), and core (external stimuli) stated.

In Feanzoni et al. [4] transfer learning is used to recognize crowd emotion. This algorithm is used in music and speech processing. Transfer learning uses a pre-trained network for low-level features and fine-tuning the network. They have a labeled data set with crowd sounds. The sound is transformed into images by encoding all the relevant sound features like spectrograms of frequency and sound features. This is Heterogeneous Transfer learning, and the flowchart of the algorithm is shown 8. First, the labeled sound is split into blocks of 2 seconds, and spectrograms are created. Spectrograms represent the short-term power spectrum of a sound and transform them into a bi-dimensional feature map with frequency and time axis to provide the amplitude. The next step is Knowledge transfer training, where these spectrograms are provided to the CNN which are pre-trained, to increase the speed of calculation and classify the crowd emotion.

In [12], proposes a model to classify emotions based on multiple levels. First, from the video, using CNN, facial emotion is extracted from all individuals, next a spatial group level ER (NVPF) is utilized, and finally, a temporal NVPF structure is built to track the frames to classify the video frames.

In [18], using 4 features from the video, the authors classify the emotion of the crowd. But these features can have some uncertainty and they suggest using fuzzy inference systems to evaluate crowd emotions. The proposed framework is shown in Fig 9. Each feature is fuzzified into 5 levels. Enthalpy and

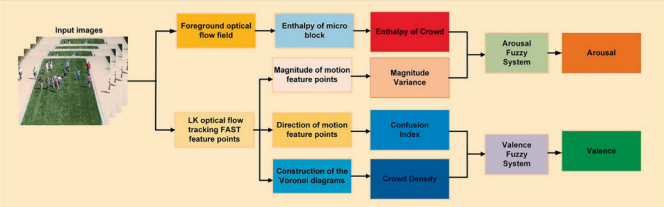

Fig. 9. Crowd emotion evaluation based on fuzzy inference of arousal and valence framework extracted from [18]

Fig. 10. Group-level Emotion Recognition on Crowd Videos extracted from [12]

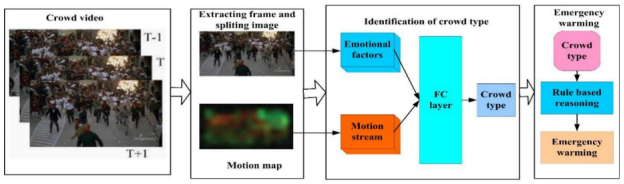

Fig. 11. DL with Crowd Emotion model flowchart extracted from [6]

Variance is used to find Arousal, whereas Confusion index and crowd density is used to determine the valence. Using 25 fuzzy inference rules a score for arousal and valence are obtained.

Deep learning methods perform well in crowd emotion recognition, as they extract relationships between features. But these models have the limitation of explainability. The use of traditional CNN is limited as it can not handle time series. Wu et al. in [17] propose a model using RNN based on a causal spatiotemporal structure to identify crowd emotions. Fig 4, shows the flowchart of the methodology used in the paper. After extracting the features from the videos, Individual Treatment Effect(ITE) module is used to obtain the causal relationship between variables and provides an explainability of the model. If the ITE between features increases above a threshold, it is provided to the RNN to pre-train the model, in order to recognize the emotion.

5) Output: In [18], the output obtained is the score for valence-arousal from which the emotions can be determined and they classify 12 emotions. In [13], classifies the video into five behavior categories, namely, 'obstacles, panic, congestion, fight and neutral.'

In [3] 3 emotions are detected using sound clips. the categories are approval, disapproval and neutral, whereas in [4], the emotions detected are joy, anger, and neutral. In [12], the output from this model is for, each video or image, and a multi-level emotion classification is provided. The 3 categories are individual faces, group-level, and entire video, as shown in 10. Individual faces are recognized into 8 categories, whereas for group level and entire video, 3 emotions are identified.

In [8], participants post their photos on to an application, and every fifteen minutes based on the emotions detected from the participants, the Christmas tree lights are changed as shown in Fig3. In [17], the model is able to classify emotions into 6 categories: angry, sad, excited, afraid, happy, and neutral.

In the rest of the models, they classify the behavior of the group, by identifying the emotions felt by the crowd. In [7], [6], after feature extraction from the video, and training the model using deep learning algorithms, the output is to predict the behavior of the crowd. It classifies them in one of the multi-class behavior. This prediction can be useful in warning the forces, to avoid violent actions. The flowchart used is shown in Fig 11 In [10], suggests image processing steps that extract edges and emotions from the chaotic scenes in indoor events. These videos result in an occluded view, but the algorithm provided in this research produced promising results to accurately classify the behavior.

6) Evaluation: In order to evaluate the models presented in the literature review, the researches performed evaluations. The evaluation was performed by using available datasets, or own curated datasets. With these datasets, using various metrics the models' performance was calculated. In [7], [6], AUC(Area Under the Curve) and ERR (Equal Error rate) were calculated. The accuracy from confusion matrix was calculated and using their algorithm their model obtained better rates compared to baseline analysis.

A dataset of images was created to evaluate the model proposed by Martini in [8]. They found the emotions angry and surprise had poor accuracy rates. Rabiee et al in [13], curated their own dataset, by annotating, the emotions present in the videos. Using low-level feature extraction as baseline analysis, they found by considering emotions, the prediction of the behavior of the crowd was more accurate. Faiscal et al in[3] curated a sound dataset. They used F1 score to evaluate their model and found that disapproval emotion was poorly identified. Patwardhan et al in [10], creates own dataset with annotated emotions. They used videos from crowded indoor events. By splitting the dataset to train and test set 70-30% they calculated the model's accuracy and recall rate. Accuracy for the emotion 'Happy' was high, whereas 'Anger' was least. The main reason was the participants had high movements, and sometimes moved away from the capture spot or covered their heads with pillows or helmets. Tripathi et al. in [16], aimed to detect protests in any given image. Since protests are allowed by all governments, classifying the emotions in protests is key. To evaluate their model, they chose accuracy as their evaluation metric and obtained a good rate. Baig et al in [1], evaluated their model, by curating a dataset for realistic crowd behavior. Different scenarios were simulated in a virtual world with different situations which are Herding, Turbulent, Stop-and-go, and Normal scenarios. They generated 3 negative and 1 normal scenario. Using this simulation, they measure the crowd's emotions and obtained fair results. In one of the negative scenarios, herding behavior, the accuracy dropped,

since the crowd became immobile and the trajectory of the crowd reduced, giving rise to less accuracy. In Franzoni et al. [4], to build the data set of sound clips, they collected their data set from cheering, rioting, and neutral crowd sounds from YouTube videos. The videos helped them to label the sound clips correctly. They used 80% of the clips for training the model and 20% to test. They randomly selected the training and test clips, which lead to overfitting and an increase in accuracy. Hence, they evaluated using different clips in training and testing. They measured their model using accuracy and obtained good results. The model proposed by Quach et al in [12], is evaluated with the self-curated datasets and other public datasets. They used accuracy as their evaluation metric. The experiments showed robust and effectiveness of each component included in their model.

Wu et al in [17], use MEDataset to evaluate their model using accuracy metric. They compared their model CS-RNN without ITE and found that ITE helped to improve their accuracy. Their approach performed better on emotions angry, sad, and neutral. Happy and excited had poor performance, since their model used motion features and both these two emotions are considered similar.

Zhang et al. in [18] use Mean Absolute Error(MAE) and Mean Relative Error to evaluate the performance of the model. They curate their own dataset based on publicly available video surveillance datasets. They obtained low MAE and MRE values showing their models performed better.

IV. Induce emotions

Reviewing the state-of-the-art research on crowd emotions models, we want to explore if there are any tools to ”induce” these emotions for enhancing digital experience. For example, if we want to watch a concert, can we feel the emotions felt during a concert from the comfort of our home. A review of our Q3: Can crowd emotions be induced during digital experiences?

In this section, we summarize how affective haptic devices(AHD) can be utilised to induce emotions. These devices are wearable technology that provides haptic feedback by vibrating and applying pressure to induce emotions. This can allow the wearer of the device to experience the emotions in real time. For instance, in [9] audience of a movie theatre wear haptic devices to map emotion of music to help hearing-imparied feel the emotions in the movie. They found that using haptic signals with high intensity and high frequency, they could produce high valence and arousal.

Jiao et al in [5], analyses current affective haptics based on emotional feedback characteristics. They present a variety of AHD and haptic signals to express emotional belongings.

In Soderstrom et al. [15], performed an evaluation on the effect of haptic feedback by using haptic vests and haptic feedback on game controllers. The results show that the haptic feedback made them more immersed in the game. Participants of the evaluation said that they were aware of everything around them, and could feel it when they got shot while wearing a haptic vest.

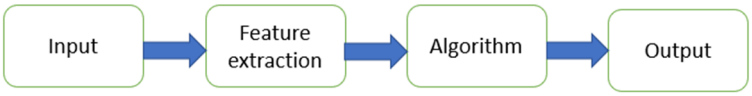

Fig. 12. Pipeline for crowd emotion recognition

The study by Peng et al. in [11] aimed to investigate the feasibility of using bio-signal haptic feedback to transmit emotions experienced by a fan to non-fan participants during a Formula 1 racing event. They found that the non-fan participants could feel the excitement of the fan participants through the bio-signal haptic feedback.

The use of affective haptic devices alone to induce emotions has limitations because emotions may not be elicited the same way for all participants using just touch sensation. Different people can experience touch differently, and their emotional response to touch can also be influenced by various factors such as culture, previous experiences, and individual differences in sensitivity to touch. Therefore, incorporating visual and sound effects can enhance the emotional experience and make it more consistent across participants.

V. Discussion

To answer our Q1, pipeline followed in crowd emotion detection model, we observe that there are four key elements needed as shown in Fig 12. Input modalities include the data, such as video or audio. Feature extraction: involves extracting the relevant features from the input data. Algorithms contain the machine learning models such as CNN, SVM, RNN to classify the emotion. Algorithms also contain an emotion model to detect emotions in the crowd. Finally, output, is the response to the crowd emotion detected. .

To answer Q2 of where crowd emotion recognition models are utilised, we saw that they are utilized in various contexts such as shopping malls, entertainment, sports events, and public safety. For instance, in shopping malls, the model’s output can be used to determine the emotions of the crowd and adjust factors such as music and lighting to create a more joyful experience and increase revenue. In sports events, it is used to measure the emotions of the crowd, and in public safety, it helps detect potential security threats by analyzing the emotions of individuals in public places.

The models reviewed used input modalities of video or sound, and none used both together. We believe by using both modalities could be useful in making better predictions. Some of the models could not classify all the emotions. The datasets curated by these models are annotated by humans and each annotator could understand the emotions differently based on their cultural differences. In [13], we observe they are unable to clearly classify happy and excited, mainly due to annotation differences.

Li et al in [7], the mapping of emotion to behavior as shown in Table II, an emotion ’nothing’ is mapped as abnormal objectives, which the authors do not discuss further, but with

more behavior characteristics mapped to emotions the table can be extended. In [4] they consider only each block of sound clip to classify and do not consider the sequence of sounds to identify the emotions. They evaluated their model in a final world cup match and found that their model obtained only an accuracy of 0.53. Hence, considering the sequence of the sounds is key which is not applied in their model. In [16], to identify the emotions present in protests, they use only images, but this can be extended by including videos and by segmenting the video along with an attention layer to have a feedback loop to monitor the frames in the video, to identify the evolution of an incident.

After reviewing the papers, we discovered that privacy is not considered in these monitoring systems. How can we preserve the privacy of individuals when the video is shared? How is the video log saved? So it is important for the designers of crowd emotion models, to ask for consent from the participants whenever possible. Which could be a challenge. In this case, data should be destroyed after analysis or the data needs to be stored securely. Crowd emotion recognition models require extensive training and are computationally expensive. They use deep learning technology to detect emotions and need to be trained to detect various anomalies. Crowd emotion recognition models need to be context-aware, using the model in all situations cannot be useful, for example, children running chaotically in a playground is not the same in a crowded event.

To answer our question Q3, we reviewed some tools to induce these crowd emotions while having a digital experience. We propose using Affective Haptic devices to induce crowd emotions. Affective haptic devices with haptic signals along with visual and acoustic stimuli, can bring about a digital experience from home. What we should consider while using these haptic devices is, "we risk losing control not just of our data or privacy, but of our bodies themselves" according to Morelli in [15]. It is crucial to address these concerns to ensure that the use of haptic technology is safe and responsible.

VI. Conclusion

This paper compares existing crowd emotion recognition models. These models are being used in various contexts to monitor the emotions of the crowd. They can bring a joyful shopping experience and protect the public from fatal accidents. These models follow a similar pipeline to build them and they are presented in this study. We have also reviewed affective haptic devices to find if they can be used to have digital experiences of crowd events from home.

This literature review is not extensive as not all state-of-the-art crowd emotion techniques are considered in our study. Although our paper does not consider all the criteria present to design a crowd detection model, designers of these models can use this paper as a general overview to build their pipeline. In future work, more criteria can be added to build a comprehensive model.

References

[1] Mirza Waqar Baig, Emilia I Barakova, Lucio Marcenaro, Carlo S Regazzoni, and Matthias Rauterberg. Bio-inspired probabilistic model for crowd emotion detection. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 3966–3973. IEEE, 2014.

[2] Iris Bakker, Theo Van Der Voordt, Peter Vink, and Jan De Boon. Pleasure, arousal, dominance: Mehrabian and russell revisited. *Current Psychology*, 33(3):405–421, 2014.

[3] Md Ahasan Atick Faisal, Mosabber Uddin Ahmed, and Md Atiqur Rahman Ahad. Eslce: A dataset of emotional sounds from large crowd events. In *2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pages 1–7. IEEE, 2021.

[4] Valentina Franzoni, Giulio Biondi, and Alfredo Milani. Crowd emotional sounds: spectrogram-based analysis using convolutional neural network. In *SAT@ SMC*, pages 32–36, 2019.

[5] Yang Jiao, Yingqing Xu, and Xiaobo Lu. Affective haptics research and interaction design. In *International Conference on Human-Computer Interaction*, pages 134–143. Springer, 2019.

[6] Xiao Li, Yu Yang, Linyang Li, and Yiming Xu. Using deep learning models combined with crowd emotion models to identify abnormal behaviors in crowds. In *Journal of Physics: Conference Series*, volume 1622, page 012051. IOP Publishing, 2020.

[7] Xiao Li, Yu Yang, Yiming Xu, Chao Wang, and Linyang Li. Crowd abnormal behavior detection combining movement and emotion descriptors. In *Proceedings of the 2nd International Conference on Industrial Control Network And System Engineering Research*, pages 106–110, 2020.

[8] Massimo Martini, Andrea Felicetti, Marco Mameli, Raffaele Vaira, Rocco Pietrini, Salvatore La Porta, Fabrizio Marconi, and Isabella Lazzini. Crowd emotion detection to light up a smart christmas tree. In *2019 IEEE 23rd International Symposium on Consumer Technologies (ISCT)*, pages 34–36. IEEE, 2019.

[9] Antonella Mazzoni and Nick Bryan-Kinns. How does it feel like? an exploratory study of a prototype system to convey emotion through haptic wearable devices. In *2015 7th International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN)*, pages 64–68, 2015.

[10] Amol Patwardhan. Edge based grid super-imposition for crowd emotion recognition. *arXiv preprint arXiv:1610.05566*, 2016.

[11] Shimeng Peng. Excitement projector: Augmenting excitement-perception and arousal through bio-signal-based haptic feedback in remote-sport watching. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–6, 2022.

[12] Kha Gia Quach, Ngan Le, Chi Nhan Duong, Ibsa Jalata, Kaushik Roy, and Khoa Luu. Non-volume preserving-based fusion to group-level emotion recognition on crowd videos. *Pattern Recognition*, 128:108646, 2022.

[13] Hamidreza Rabiee, Javad Haddadnia, Hossein Mousavi, Moin Nabi, Vittorio Murino, and Nicu Sebe. Emotion-based crowd representation for abnormality detection. *arXiv preprint arXiv:1607.07646*, 2016.

[14] Francisco Luque Sánchez, Isabelle Hupont, Siham Tabik, and Francisco Herrera. Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. *Information Fusion*, 64:318–335, 2020.

[15] Ulrik SöDerströM, William Larsson, Max Lundqvist, Ole Norberg, Mattias Andersson, and Thomas Mejtoft. Haptic feedback in first person shooter video games. In *Proceedings of the 33rd European Conference on Cognitive Ergonomics*, pages 1–6, 2022.

[16] Gaurav Tripathi, Kuldeep Singh, and Dinesh Kumar Vishwakarma. Crowd emotion analysis using 2d convnets. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 969–974. IEEE, 2020.

[17] Minzhong Wu, Lei Wang, and Guodong Li. Crowd emotion recognition based on causal spatiotemporal structure. In *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*, pages 368–374, 2022.

[18] Xuguang Zhang, Xiuxin Yang, Weiguang Zhang, Gongfa Li, and Hui Yu. Crowd emotion evaluation based on fuzzy inference of arousal and valence. *Neurocomputing*, 445:194–205, 2021.